Data Mining Metrics

Himadri Barman

Data Mining has emerged at the confluence of artificial intelligence, statistics, and databases as a technique for automatically discovering summary knowledge in large datasets. Data mining first requires understanding the data available, developing questions to test, and finally drawing conclusions from data analytic results. Metrics are some parameters or measures of quantitative assessment used for measurement or comparison in a given context. A metric for all practical purpose is just a variable. It needs to be clearly defined. The number of metrics needs to be kept under control to ensure that the measuring task is achievable. It is thus reasonable to expect that as the context changes, the metrics would change. Literature has not defined Data mining metrics as such. Data mining metrics may be defined as a set of measurements which can help in determining the efficacy of a Data mining Method / Technique or Algorithm. They are important to help take the right decision as like as choosing the right data mining technique or algorithm.

Data mining comes in two forms. Directed data mining involves searching through historical records to find patterns that explain a particular outcome and includes the tasks of classification, estimation, prediction and profiling. Undirected data mining searches through the same records for interesting patterns. It includes the task of clustering, finding association rules and description. Data mining models are the key for both. Each type of model so designed will have its own metrics by which it can be assessed, but there may be assessment tools that are independent of the type of model. In many cases, a single metric may not be sufficient to evaluate. In such cases, we might have to look at multiple metrics which can be used to validate one another and maximize the accuracy of the evaluation. Choosing the right metrics for the assessment is of paramount importance.

Data mining metrics generally fall into the categories of accuracy, reliability, and usefulness. Accuracy is a measure of how well the model correlates an outcome with the attributes in the data that has been provided. There are various measures of accuracy, but all measures of accuracy are dependent on the data that is used. In reality, values might be missing or approximate, or the data might have been changed by multiple processes. Particularly in the phase of exploration and development, we might decide to accept a certain amount of error in the data, especially if the data is fairly uniform in its characteristics. For example, a model that predicts sales for a particular store based on past sales can be strongly correlated and very accurate, even if that store consistently used the wrong accounting method. Therefore, measurements of accuracy must be balanced by assessments of reliability.

Reliability assesses the way that a data mining model performs on different data sets. A data mining model is reliable if it generates the same type of predictions or finds the same general kinds of patterns regardless of the test data that is supplied. For example, the model that we generate for the store that used the wrong accounting method would not generalize well to other stores, and therefore would not be reliable.

Usefulness includes various metrics that tell us whether the model provides useful information. For example, a data mining model that correlates store location with sales might be both accurate and reliable, but might not be useful, because you cannot generalize that result by adding more stores at the same location. Moreover, it does not answer the fundamental business question of why certain locations have more sales. We might also find that a model that appears successful in fact is meaningless, because it is based on cross-correlations in the data.

Measuring the effectiveness or usefulness of data mining approach is not always straightforward. In fact, different metrics could be used for different techniques and also based on the interest level. From an overall business or usefulness perspective, a measure such as Return on Investment (ROI) could be used. ROI examines the difference between what the data mining technique costs and what the savings or benefits from its use are. Of course, this would be difficult to measure because the return is hard to quantify. It could be measured as increased sales, reduced advertising expenditure, or both. In a specific advertising campaign implemented via targeted catalog mailings, the percentage of catalog recipients and the amount of purchase per recipient would provide one means to measure the effectiveness of the mailings.

There can be a more computer science / database perspective to measure various data mining approaches. It is assumed that the business management has determined that a particular data mining application be made. They subsequently will determine the overall effectiveness of the approach using some ROI (or related – like TCO: Total Cost of Ownership) strategy. The objective then is to compare different alternatives to implementing a specific data mining task. The metrics used include the traditional metrics of space and time based on complexity analysis. In some cases, such as accuracy in classification, more specific metrics targeted to a data mining task may be used.

Evaluation metrics play a critical role in data mining. Metrics are used to guide the data mining algorithms and to evaluate the results of data mining. For example, when using a decision tree algorithm to solve a classification task, information gain may be used to guide the construction of the decision tree while accuracy may be used to evaluate the performance of the final tree.

The development of a large number of rule induction and decision tree construction algorithms for data mining by researchers in machine learning and statistics has seen empirical evaluation and justification become an important aspect for acceptance of newly developed algorithms by researchers in the field. To provide a comprehensive evaluation, a set of standard criteria is needed such as: induction time, size of induction results, time to execute the induction results, and predicative accuracy. One algorithm may be able to perform better than others with one criterion, but may perform poorly with other criteria. With the same set of algorithms, we can also get different evaluation results from different sets of databases. The question of why, and under which circumstances one algorithm (whether it is newly designed or an existing one) outperforms others becomes more important than simply presenting empirical results from an arbitrarily selected set of databases. Research on data mining metrics is based on the above mentioned, widely adopted criteria. These metrics also look into the characteristics of the data sets for experiments such as: the numbers of classes, attributes and examples, the distribution of training examples in the example space, the level of noise and the mixture of continuous and nominal values. The aim is to develop a meaningful set of metrics with well documented experiment results for different algorithms. These metrics can be used as a test bed for newly developed algorithms against existing ones. There is now lot of intent in developing and designing data-mining metrics that can be used to make predictive models that support systemic change.

Data mining has now become specialized like those on web data (web mining), spatial data, etc. With the explosion in web generated data, web mining has found many takers. There are many web mining metrics, like website visitors, pages served, indegree or queries in a given time, etc. that can be tracked. Data mining on spatial data has become important due to the fact that there are huge volumes of spatial data now available holding a wealth of valuable information. Distance metrics are used to find similar data objects that lead to develop robust algorithms for the data mining functionalities such as classification and clustering.

Many modern businesses are data driven. A great deal of effort is spent on using masses of data to guide decisions at all levels. When data mining algorithms are transferred into the business community, the technical metrics associated with the algorithm are also transferred. Practitioners in the business world are then able to evaluate predictive business models developed with the available technical metrics. Therefore, at the core of these efforts are metrics. Businesses thus focus on producing timely, correct and relevant metrics that help them in their operations. Data mining metrics should be directly proportional to the improvement in the data mining operations. Since, a large number of data mining metrics are there, they should be selected with caution. The data mining metrics needs to be clearly defined and avoid any kind of ambiguity in interpretation. It is believed that data mining metrics should be flexible enough to meet changing needs and requirements. In an interesting conclusion, it is worthwhile to mention that data mining metrics has become a niche field with many top IT consultants to give advices/suggestions. It can become a career for many!

References:

- Data Mining: Introductory and Advanced Topics by Margaret H Dunham
- Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management by Michael J. Berry, Gordon S. Linof
- Fast Distance Metric Based Data Mining Techniques Using P-trees: k-Nearest-Neighbor Classification and k-Clustering – A Thesis submitted by Md Abdul Maleq Khan
- CFP: A Special Issue of Informatica On Data Mining Metrics – a forward note by Dr. X D Wu
- Mining with Rarity: A Unifying Framework by Gary M. Weiss
- Knowledge Discovery and Data Mining: Challenges and Realities by Xingquan Zhu, Ian Davidson
- http://msdn.microsoft.com/en-us/library/ms174493.aspx accessed on September 10, 2012 at 0825 hrs